图书情报工作
第65卷 第14期 2021年7月

基于问题－方法组合的科技论文新颖性度量与创新类型识别*

■ 钱佳佳　罗卓然　陆伟
武汉大学信息管理学院　武汉 430072

摘　要：[目的/意义]科技论文的新颖性度量是科技成果评价的重要内容，本文旨在从科技论文的核心要素即问题和方法出发，提出一种基于问题－方法组合的科技论文新颖性度量与创新类型识别方法。[方法/过程]基于词频原则分别计算科技论文的问题新颖度、方法新颖度、问题－方法组合新颖度，再通过权重赋值计算论文整体的新颖度。同时，基于组合创新理论，从科技论文问题－方法组合的角度出发提出4种创新类型以及根据文章新颖值判断其所属创新类型的方法。[结果/结论]对1951－2018年的20多万篇ACM论文进行实证研究，证明提出的科技论文新颖性度量方法以及创新类别识别方法是科学、合理和可操作的。

关键词：科技论文　新颖性度量　组合创新　问题－方法组合
分类号：G250
DOI：10.13266/j.issn.0252－3116.2021.14.010

1 引言

“民族进步的灵魂”，习近平总书记这样谈创新。《中共中央关于制定国民经济和社会发展第十四个五年规划和二〇三五年远景目标的建议》[1]将“坚持创新”列为未来五年十二项重要领域工作的首位，在科技创新上强调要强化国家战略科技力量，激发人才创新活力，完善科技创新体制机制，坚决破除“唯学历、唯职称、唯学历、唯奖项”，科技成果创新评价成为重要任务。科技论文是科技成果的主要载体，具备创新性的论文才能对科学发展有所贡献，科技论文的创新评价对科技评价、科研经费调配等都具有很高的价值。

科技成果的创新性在科技查新中体现为在查询委托日以前查新项目的科学技术内容部分或者全部没有在国内外出版物上公开发表过[2]；在申请发明和实用新型专利时体现为申请专利的发明或者实用新型必须不同于现有技术[3]；在学术论文上，陈建青[4]等将“创新”定义为在相关学术领域内，创立或发展了有价值的新理论、新方法、新技术等，或在前人研究成果的基础上，加工、整理、发掘出新意，如提出新结论等，可见新颖性是创新性的关键特征，创新性评价离不开新颖性评价。

创新成果通常不是凭空产生的，组合是创新的核心之一，有不少学者的研究论证了这一点：I. Nonaka[5]认为组合是组织运用显性知识进行知识创造的主要途径；L. Fleming[6]认为已有知识的重新组合和新知识的组合都有可能带来创新；B. Uzzi等[7]提出不论在哪个学科领域，现有知识的非典型组合可能带来创新；S. Mishra等[8]在探索MeSH词和论文新颖性的关系时提出，在生物医学中主题词的组合能够反映文章的新颖性，影响最大的文章往往是在典型组合的基础上引入了一些新颖的组合。可见，组合新颖性作为科技论文新颖性的计量方法已经具备一定的理论基础。问题和方法是科技成果的两大核心要素，每篇科技论文都是新老问题和新老方法的交叉组合，问题和方法的重新组合和新问题新方法的组合都有可能带来创新，因此计算问题－方法的组合新颖性能够在一定程度上衡量学术论文的新颖性。

受限于学术论文的问题和方法难以提取，以往研究较少有从科技论文的问题、方法出发计量其新颖性，

* 本文系国家社会科学基金重大项目“基于认知计算的学术论文评价理论与方法研究”（项目编号：17ZDA292）研究成果之一。
作者简介：钱佳佳（ORCID：0000－0002－6058－1287），硕士研究生，E-mail：jiajiaqian@whu.edu.cn；罗卓然（ORCID：0000－0003－0677－8350），博士研究生；陆伟（ORCID：0000－0002－0929－7416），教授，博士生导师。
收稿日期：2021－01－11　修回日期：2021－03－01　本文起止页码：82－89　本文责任编辑：徐健

而是不区分语义功能地计算关键词词频或相似性,结果会存在一定的偏差,如某种技术方法早期被作为研究对象加以研究,成熟后被应用于其他问题,其应用的文章也具有很高的新颖性,但如果不加区分地计算词频或者相似性,其新颖性会受早期作为研究对象时的文章影响而变得较低。陆伟等以从《计算机学报》《情报学报》等多本计算机及图书情报领域期刊获得的2009－2018年刊载的12多万篇文献为数据基础,构造了一种基于规则的数据标注方法对数据进行标注,并用BERT预训练模型对输入的文本进行向量化表征,利用长短期记忆网络模型(Long Short-Term Memory,LSTM)对关键词进行自动判别以实现论文的问题、方法的识别[9]。基于该研究成果,本文提出了一种基于问题－方法组合共现率度量科技论文新颖性的方法。此外,根据组合创新理论,提出了基于科技论文问题－方法组合的4种创新类型以及依据文章新颖值识别其所属创新类型的方法。

2 相关研究

2.1 科技论文新颖性度量

目前已有以下几类评价单篇科技论新颖性的方法:

第一种是同行评议法,是学术界最为通用的一种主观定性评价方法,依靠领域评审专家个人认知进行评价,操作简单易行,但实践中易产生因个人认知特性造成的非公正性、非客观性和非合理性等问题[10]。

第二种是基于引用关系计算学术成果的新颖性,理论基础是学术成果的影响力体现在被其他成果引用中,有学者探讨了文章被引量和新颖性之间的关系,例如有沈律[11]认为科技成果的创新性与引用率成正比,即科技论文引用率越高,其创新度越高。逯万辉等[12]对国内图书情报领域期刊论文的新颖性和被引量进行统计分析,发现学术成果主题新颖性与学术引用之间存在显著的正相关关系,主题新颖性较高的学术成果被引情况高于新颖性较低的学术成果。虽然文章的被引情况能在一定程度上反映论文的新颖性特征,但引文分析法的局限在于从科技成果外在特征评价其新颖性,没有深入到文本层面去度量内容的新颖性。

第三种是从学术论文本身出发,基于论文关键词词频或相似性来测度学术文本主题新颖性。杨建林等[13]吸收词频原则、逆文档频率原则等提出了带时间戳的关键词对逆文档频率以量化文档主题的新颖性,发现同一学科领域中重要核心期刊刊载论文的平均主题新颖度要高于普通期刊。任海英等[14]基于主题词共现网络计算文章的新颖组合率,当新颖组合率高于一定阈值时,就认为该文章具有创新性。杨京等[15]利用Jaccard系数计算文章关键词的重叠度,认为关键词重叠度越高,文章间相似度越大,则主题新颖性越低。许丹等[16]抽取文章中的主题词,计算其逆文档频率从而计算文章的主题新颖性。逯万辉等[17]基于Doc2Vec和隐马尔科夫模型(Hidden Markov Model, HMM)算法计算文本相似度,从而计算学术成果主题新颖性。钱玲飞等[18]认为学术论文的学科交叉程度越高,其创新度可能越大,基于关键词词频定义关键词交叉率等指标量化学科的创新能力;此外,其还定义了共现关键词的生命指数和有效新词出现率,以比较学科创新力,发现有效新词出现率越高则学科创新保持力越强。

虽然已有的方法可以在一定程度上评价学术论文的新颖性,但在计算关键词词频或者相似度的时候并没有考虑关键词的语义功能信息,如"genetic programming"在"*Multi-chromosomal genetic programming (2005)*"一文中是研究问题,而在"*Genetic programming for shader simplification (2011)*"一文中是方法,不加区分地计算词频或相似性会将问题和方法混合,导致后一篇文章的新颖值受前者的影响降低,但实际上其作为一项新技术应用到新老问题上应具有高新颖性。本文基于深度学习模型得到科技论文的问题词和方法词,主要通过计算问题－方法组合的共现率(即问题－方法组合出现的频率)衡量科技论文的新颖度,计算过程严格区分了问题词和方法词,避免了语义功能不同的词的新颖值相互影响的问题。此外,与以往研究不同的是,本文在计算问题－方法词对出现频率的同时,还计算了单个问题、方法出现的频率。之所以这样计算,是因为新问题＋新方法、新问题＋老方法、老问题＋新方法作为组合出现的频率都是0,仅考虑组合的共现率,其新颖值都是1,但通常认为新问题＋新方法组合的新颖性要高于新问题＋老方法和老问题＋老方法的组合,并且组合中老问题或者老方法出现的频率越高,其新颖性越低。

2.2 科技论文创新类型识别

创新的概念最早由经济学家熊彼特于1912年在《经济发展理论》中提出[19],之后得到不断地研究和深化。关于创新的分类,有学者根据创新的大小进行划分,如R. Garcia等将创新分为根本型创新、适度型创新和渐进型创新[20];也有学者从知识管理的角度划分创新类型,如R. M. Henderson等将创新分为渐进型创

新、构建型创新、模型型创新和根本型创新,同时认为创新活动所运用的新知识可能强化现有知识也有可能摧毁现有知识[21];根据创新所依赖的价值网络(市场)的不同,J. L. Bower 等将创新分为延续型创新和破坏型创新[22]。从创新内容出发,宋子良将创新分为理论创新、方法创新和交叉创新[23]等。除此之外,还有学者从学科、形态、重要性、用途等角度划分创新类型。

尽管已有的创新分类方法不少,但是大多数研究停留在理论提出层面,鲜有学者提出具体的、可操作的创新类型识别方法。基于此,本文从科技论文问题-方法组合的角度出发提出 4 种创新类型,并提出一种依据论文新颖值计算结果识别其所属创新类别的方法。

3 基于问题-方法组合的科技论文新颖性度量与创新类型识别

3.1 基于问题-方法组合共现率的科技论文新颖性度量

单篇科技论文的问题-方法词对是新老问题和新老方法的交叉组合,具体包括新问题+新方法、新问题+老方法、老问题+新方法和老问题+老方法(如图 1 所示)。从问题、方法出发评价学术论文的新颖性,一般认为包含新问题+新方法的文章新颖性最高,其次是包含新问题+老方法或者老问题+新方法的文章,最低是包含老问题+老方法的文章。老问题+老方法的组合并不意味着论文没有新颖性,比如老问题+老方法的新组合也具有较高的新颖性。同时某问题-方法对所在的文章发表得越早,其新颖性越高。总之,科技论文的问题-方法组合能够在一定程度上体现文章新颖性的高低。

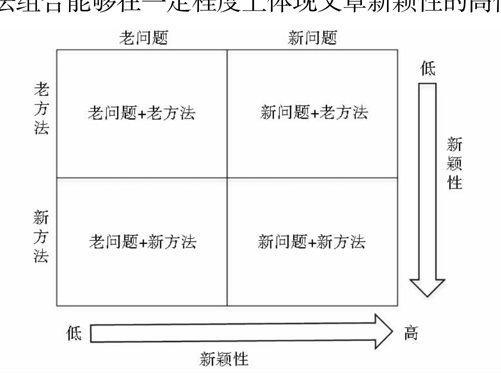


图 1 学术论文的问题-方法组合及新颖性高低

学术论文中出现频率较高的问题、方法词往往能够反映领域的研究热点,而出现频率较低的问题或方法则可能反映了论文的新颖之处,比如新问题的提出或者新方法的应用。学术论文的新颖性会随着其问题、方法在该文献发表前出现频率的增加而降低。基于此,本文引入词频原则度量学术论文新颖值。问题-方法组合中存在两种词频,一种是问题-方法词作为组合出现的频率,即组合共现率;另一种是问题词、方法词作为单个词出现的频率。某科技文章发表前,文章中的问题-方法组合在该领域已发表论文中出现的频率越高,说明问题-方法组合的新颖性越低,该科技论文的新颖性越低;此外,文章中单个问题或方法在该领域已发表论文中出现的频率越高,说明该问题或方法自身的新颖性越低,包含该问题或者方法的科技论文的新颖性受其影响也会变低。如果某对问题-方法词在文章发表以前出现的频率为 0,那么其是一对崭新的问题方法组合。需要强调的是,本文计算的科技论文新颖值均是与文章发表之前的领域内文章对比计算而得的,讨论的是文章发表时的新颖值,而非其他时间点上的文章新颖值。

根据学术论文问题-方法组合的新颖性规律以及词频原则,本文提出了单篇科技论文的问题新颖性、方法新颖性和问题-方法对新颖值的计算公式,如公式(1)(2)(3)所示:

$$nov(Q)=\frac{\sum_{i=1}^{|Q|}\frac{1}{\ln(n(Q_i))+1}}{|Q|} \qquad 公式(1)$$

$$nov(M)=\frac{\sum_{j=1}^{|M|}\frac{1}{\ln(n(M_j))+1}}{|M|} \qquad 公式(2)$$

$$nov(Q,M)=\frac{\sum_{i=1}^{|Q|}\sum_{j=1}^{|M|}\frac{1}{\ln(n(Q_i,M_j))+1}}{|Q|\ |W|} \qquad 公式(3)$$

其中,Q,M 分别是文档 D 的问题词集合与方法词集合,$nov(Q)$ 表示文章问题的新颖值,$nov(M)$ 表示文章方法的新颖值,$nov(Q,M)$ 表示文章问题-方法对的新颖值。$|Q|$ 表示集合 Q 的元素个数,即文档 D 的问题数量;$|M|$ 表示集合 M 的元素个数,即文档 D 的方法数量。Q_i、M_j 分别表示文档 D 中第 i 个问题词和第 j 个方法词;$n(Q_i)$ 表示截至 D 发表时,问题 Q_i 在同领域出现的频数;$n(M_i)$ 表示截至 D 发表时,方法 M_i 在同领域出现的频数;(Q_i,M_j) 表示问题-方法对,$n(Q_i,M_j)$ 表示截至 D 发表时,问题-方法对 (Q_i,M_j) 在同领域出现的频数(频数计算时包含文档 D)。取 ln 是为了减缓新颖值随频数的下降速率,避免新颖值过低,同时又能保持论文新颖值的大小顺序。

随后,本文取文章的问题新颖值、方法新颖值、问题-方法组合新颖值的加权平均值作为科技论文整体的新颖性值,记为 $nov(D)$,具体计算方法如公式(4)所

示，其中，k_1，k_2，k_3（$k_1 + k_2 + k_3 = 1, k_1, k_2, k_3 \geq 0$）分别为问题新颖性、方法新颖性和问题－方法对新颖性的权重，其取值大小能够反映文章所属领域的问题、方法和问题－方法组合对文章新颖性的决定程度，取决于文章所属领域知识更新的特征。计算不同领域文章的新颖值时应该分析领域的研究更新特征，根据具体的情况确定权重的大小。

$$nov(D) = k_1 nov(Q) + k_2 nov(M) + k_3 nov(Q,M) \quad \text{公式(4)}$$

3.2 基于问题－方法组合新颖度的科技论文创新类型识别

新颖性是学术论文创新性的本质属性，当文章的新颖值大于一定阈值时，可以认为该文章具有较高的创新性并根据其新颖值的大小划分创新类型。显然，问题、方法单个出现的次数均大于等于问题－方法作为组合出现的次数，根据公式（1）、（2）、（3）可计算得出单篇科技论文的 $nov(Q,M) \geq nov(Q)$，$nov(Q,M) \geq nov(M)$，$nov(D) \leq nov(Q,M)$，即当单篇科技文献的新颖值大于一定阈值时，其问题－方法对的新颖值也一定大于该阈值。若单篇科技论文中方法词的新颖值大于一定阈值时，可以认为其存在方法创新；若问题词的新颖值大于一定阈值，则可以认为该文章存在问题创新。综上，根据科技论文问题、方法、问题－方法对新颖值的大小能够划分其所属创新类型，具体判断如下：

$$\text{当 } nov(D) \geq T_d \text{ 时，}\begin{cases} \text{nov}(Q) \geq T_q, \text{nov}(M) \geq T_m, \text{“新问题＋新方法”类组合创新} \\ \text{nov}(Q) \geq T_q, \text{nov}(M) \leq T_m, \text{“新问题＋老方法”类组合创新} \\ \text{nov}(Q) \leq T_q, \text{nov}(M) \geq T_m, \text{“老问题＋新方法”类组合创新} \\ \text{nov}(Q) \leq T_q, \text{nov}(M) \leq T_m, \text{“老问题＋老方法”类组合创新} \end{cases}$$

其中，T_d 是文章具备创新性的阈值，T_q 是问题创新的阈值，T_m 是方法创新的阈值，图 2 展示了整个判断流程。在文章整体新颖值大于规定阈值的情况下，若问题和方法新颖值均大于对应阈值，则其为“新问题＋新方法”类组合创新；若问题大于对应阈值，方法否，则为“新问题＋老方法”类组合创新；若方法大于对应阈值，问题否，则属于“老问题＋新方法”类组合创新；若问题和方法均小于等于对应阈值，则属于“老问题＋老方法”类组合创新，该组合中问题和方法虽已单个出现多次，但作为组合是新颖的。

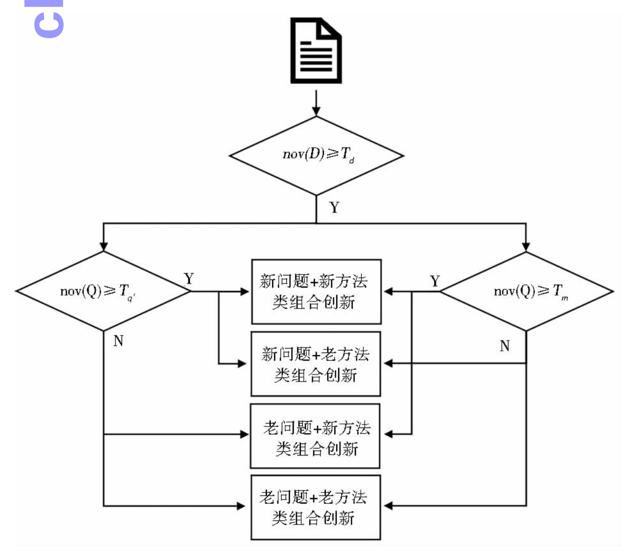


图 2　基于问题－方法组合的科技论文创新类别判断流程

4　实证研究

4.1　数据集构建

本文首先采集了 ACM（Association for Computing Machinery，美国计算机协会）从 1951 年至 2018 年间的计算机领域论文，抽取文章的标题和摘要数据，基于陆伟等提出的问题方法抽取模型识别出论文的问题词和方法词。其次将每篇文章的 DOI 号、题目、摘要、问题词、方法词、发表时间以及截至 2018 年 12 月的文章被引量存入 MySQL 数据库，数据库如图 3 所示，总计 204 310条有效数据，文章数量按年分布情况如图 4 所示。最后根据设计的公式编写程序代码计算每篇文章的新颖值并进行分析。

4.2　科技论文新颖值计算

对数据集中所有的文章进行新颖度计算，每篇文章的新颖值计算衡量的都是其发表时在领域内的新颖性大小，实验时只需将文章的问题词、方法词和问题－方法对在文章发表前出现的频率代入计算公式即可。图 5 展示了数据集中各文章的问题、方法、问题－方法对新颖值的分布情况，其中 x 轴为问题的新颖值，y 轴为方法的新颖值，z 轴为问题－方法对的新颖值。可见问题的新颖值主要分布在 0.2－0.6 区间以及 1.0，方法新颖值呈现相似分布，表明在实验数据集中问题和方法的更新速率相似。根据问题、方法新颖值的分布规律，实验将公式（4）中的问题、方法和问题－方法对的权重分别设为 0.25、0.25 和 0.5。笔者取 0.6（本文

ABC pro	123 articl	ABC	ABC art_title	ABC art_au_tags	123 art	ABC art_pub_date	ABC art_topic	ABC art_method
2800835	2,800,847	0	Senbay: smartphone-based activity capturing and sl	application;experiment :	2,015	September 07 -	a smartphone-bas	anim two-dimension ba
1718487	1,718,526	7	Evolution of two-sided markets	coalitions;economics;eq	2,010	February 04 - 0(	two-sid market evc	a general model
3093338	3,093,378	2	A CyberGIS-Jupyter Framework for Geospatial Analy	computational reproduc	2,017	July 09 - 13	data-intens and sci	advanc cyberinfrastruct
1629395	1,629,404	1	Spatial complexity of reversibly computable DAG	compilers;concurrency;c	2,009	October 11 - 16	program revers	spatial complex
1999747	1,999,795	9	WeScheme: the browser is your programming envir	compilers;computer scie	2,011	June 27 - 29	web browser	a program environ
1602165	1,602,186	10	Sensor ranking: A primitive for efficient content-bas	algorithms;communicati	2,009	April 13 - 16	web of thing searcl	sensor rank
291080	291,094	11	Making systems sensitive to the user's time and wo	adaptive systems;bayesi	1,999	January 05 - 08	system adapt	a decision-theoret appr
773184	773,200	5	Cycle therapy: a prescription for fold and unfold on	algorithms;functional lai	2,001	September 05 -	a fold oper	cyclic structur
1811212	1,811,220	24	Modeling shared cache and bus in multi-cores for ti	design;embedded and c	2,010	June 28 - 29	multi-cor system	a time analysis framewc
2072069	2,072,071	3	Measures to establish trust in internet voting	acceptance;computing /	2,011	September 26 -	internet vote syste	trust measur
2484028	2,484,157	3	Author disambiguation by hierarchical agglomerativ	clustering;clustering anc	2,013	July 28 - August	hierarch agglom cl	adapt stop criterion
2839462	2,856,532	6	Inner Garden: an Augmented Sandbox Designed fo	calm technologies;multi	2,016	February 14 - 1	augment sand	eeg
3212734	3,212,774	0	The Energy Complexity of Broadcast	broadcast;distributed al	2,018	July 23 - 27	energy-effici broac	collis detect
1368044	1,368,075	2	WebAnywhere: a screen reading interface for the w	blind users;design;huma	2,008	April 21 - 22	blind web user	webanywher
1080730	1,080,748	1	Sizing of IEEE 802.11 wireless LANs	analytical models;mac;n	2,005	September 02 -	ieee 802 11 wireles	approxim size method
2610384	2,610,419	6	Covrig: a framework for the analysis of code, test, ai	bugs and fixes;coverage	2,014	July 21 - 25	mine dynam softw	covrig
2030112	2,030,220	1	NFC+: NFC-assisted media sharing for mobile devic	communication hardwa	2,011	September 17 -	fast and secur mot	nfc
3125739	3,125,769	0	Investigation of Approach to Others for Modeling o	communication;f-format	2,017	October 17 - 20	physic interact	mutual communic neec
1378063	1,378,085	4	Scheduling real-time multi-item requests in wireless	algorithms;collaborative	2,007	September 10 -	on-demand broadc	time-crit multi-item rec
1152215	1,152,268	7	A mobile multimodal dialogue system for public trai	design;domain specific l	2,006	September 12 -	mobil broadband ii	multimod rout navig sy
223355	223,434	2	Conversational dialogue in graphical user interfaces	design;graphics input de	1,995	May 07 - 11	aldehyd	heat
337292	337,606	0	MINFLOTRANSIT: min-cost flow based transistor siz	algorithms;design;emer	2,000	June 05 - 09	fast transistor size	minflotransit
3196959	3,196,975	0	Explanations and Transparency in Collaborative Wo	collaboration;data-centr	2,018	June 10 - 15	view program	static explan
3109453	3,109,474	0	End-to-end molecular communication channels in c	cell metabolism;informa	2,017	September 27 -	reprogram biolog (	molecular communic
1923947	1,924,030	0	6th Workshop on Challenges for Parallel Computing	algorithms;design;gener	2,010	November 01 -	multi-nod distribut	commod hardwar price
2016039	2,016,062	0	A comparison of Player/Stage/Gazebo and Microsof	computing industry;des	2,011	March 24 - 26	mobil robot develc	usabl criteria

图 3　实验数据集截图(部分)

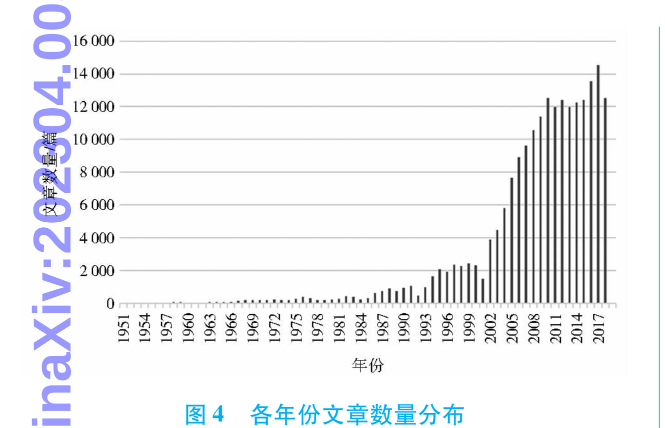


图 4　各年份文章数量分布

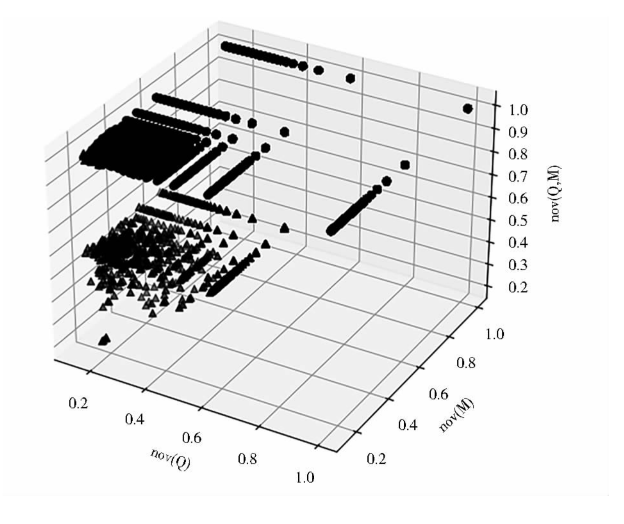


图 5　问题、方法、问题 - 方法对新颖值的散点分布

认为老问题 + 老方法的新组合也具备较高的创新性,0.6 是问题方法组合新颖值为 1 的文章的最低新颖值)为文章具备创新性的阈值进行分析,为更好地观察实验结果,实验将图 5 中文章新颖值大于等于 0.6 的文章对应的点设为圆点,其余点设为三角形,即圆点代表的是具有较高创新性的文章,三角形代表的是低创新性的文章。从图 5 可以看出,当文章整体的新颖值大于等于阈值时,文章问题 - 方法对的新颖值也大于等于该阈值,反之并不一定成立,原因是当文章的问题 - 方法组合的新颖值比较高,但是问题或方法自身出现很多次时,其文章整体的新颖值会受影响而变得较低。

当文章的新颖值小于阈值时,笔者认为其是创新性低的文章。当文章的新颖值大于一定阈值时,根据文章的问题和方法新颖值可以判断其所属的创新类型。若问题、方法的创新阈值均设为 1.0(1.0 为问题、方法新颖值的中位数),则可将新颖性较高的文章划分为新问题 + 新方法、老问题 + 新方法、老问题 + 老方法和新问题 + 老方法 4 类创新,根据统计,其中占比最大的是新问题 + 新方法创新,占比 49.22%,其次是老问题 + 新方法占比 21.72%,新问题 + 老方法占比 20.93%,老问题 + 老方法的组合创新最少,占比 8.13%。笔者认为产生上述结果主要是因为实验选用的 ACM 数据集中大多数文章都处在领域前沿,均有问题或方法创新,文章的新颖值普遍较高。

4.3　实例分析

本研究实验数据量共计 20 多万条,很难通过对每篇科技论文进行分析验证新颖值计算结果的合理性,但在科技论文新颖值计算后可以将论文划分为新颖值

较高的新问题＋新方法、新问题＋老方法、老问题＋新方法、老问题＋老方法组合的文章以及新颖值低于阈值的低创新性文章，于是笔者从上述5类文章中分别随机抽取了2篇(共10篇)进行实例分析，以证明本文提出的新颖度计算方法和创新类别判定方法的合理性，抽取出的文章信息如表1所示：

表1 随机抽取的科技论文新颖值计算及创新类型判断结果

标题	被引量	发表日期	问题词	方法词	nov (Q)	nov (M)	nov (Q,M)	nov (D)	创新类型
R-trees: a dynamic index structure for spatial searching	4808	1984－06－18	spatial searching	R-trees	1.0	1.0	1.0	1.0	新问题＋新方法
XGBoost: A Scalable Tree Boosting System	3165	2016－08－13	end-to-end tree boost	xgboost	1.0	1.0	1.0	1.0	新问题＋新方法
Cross-domain sentiment classification via spectral feature alignment	97	2010－04－26	cross-domain sentiment classification	spectral feature alignment	0.47	1.0	1.0	0.87	老问题＋新方法
RAP: an associative processor for data base management	19	1975－05－19	data base management	pointer mechanisms	0.38	1.0	1.0	0.85	老问题＋新方法
Invetter: Locating Insecure Input Validations in Android Services	0	2018－10－15	android input validation	Machine learning	1.0	0.144	1.0	0.78	新问题＋老方法
A Genetic Algorithm-Based Solver for Very Large Jigsaw Puzzles	2	2014－07－12	jigsaw puzzle solver	genetic algorithm	1.0	0.16	1.0	0.79	新问题＋老方法
Experiments with Convolutional Neural Network Models for Answer Selection	4	2017－08－07	question answer	Convolutional neural network	0.19	0.20	1.0	0.60	老问题＋老方法
What do concurrency developers ask about?: a large-scale study using stack overflow	12	2018－10－11	concurrency program	topic modeling	0.19	0.20	1.0	0.60	老问题＋老方法
Matrix and Tensor Decomposition in Recommender Systems	0	2016－09－15	recommendation system	matrix and tensor decomposition	0.16	0.21	0.47	0.33	低创新
Automatically evolving difficult benchmark feature selection datasets with genetic programming	0	2018－07－15	feature select	genetic programming	0.20	0.15	0.47	0.32	低创新

随机抽取出的文章中，属于"新问题＋新方法"类组合创新的是 *R-trees: a dynamic index structure for spatial searching* 和 *XGBoost: A Scalable Tree Boosting System*，前者首次提出了一种处理高维空间存储问题的数据结构；后者首次提出了一种可扩展的端到端基于树的 boosting 系统，这两篇文章的研究问题和方法在发表时都非常新颖，发表后都得到了大量的引用和应用。*Cross-domain sentiment classification via spectral feature alignment* 和 *RAP: an associative processor for data base management* 是属于"老问题＋新方法"类组合创新，前者提出了一种频谱特征对齐算法进行跨领域情感分类；后者提出了数据库中的 RAP——联动处理器，这两篇科技论文的研究问题分别是跨领域情感分类和数据库设计，在文章发表时都已经有一定的文献积累，但方法新颖，文章整体的新颖值较高。*Invetter: Locating Insecure Input Validations in Android Services* 和 *A Genetic Algorithm-Based Solver for Very Large Jigsaw Puzzles* 则是"新问题＋老方法"类组合创新，前一篇提出了一个叫做 Invetter 的工具，其利用机器学习的方法实现了在 Android 服务中查找不安全的输入验证；后一篇提出了一个基于遗传算法的有效自动化拼图难题求解器，两篇文章使用的方法都不是非常新颖，但是都解决了一个新的、有意思的问题，所以具有较高的新颖性。*Experiments with Convolutional Neural Network Models for Answer Selection* 和 *What do concurrency developers ask about?: a large-scale study using stack overflow* 都属于"老问题＋老方法"新组合的创新，前者是关于卷积神经网络在自动问答中的应用；后者研究的是开发者在并发编程中的问题，方法使用的是 LDA 主题模型，两篇文章发表时研究同样问题或方法的文章都有一定的积累，但是问题和方法的组合是新颖的。*Matrix and Tensor Decomposition in Recommender Systems* 和 *Automatically evolving difficult benchmark feature selection datasets with genetic programming* 是新颖值相对较低的文章，前者关于推荐系统中的矩阵与张量分解问题，后者是关于使用遗传编程自动扩充现有数据集以便更科学地测试特征选择性能的文章，问题和方法、以及问题－方法的组合都已出现过多次，故新颖性最低。综上，本文从

案例分析的角度证明了所提科技论文新颖性度量方法和创新类型识别方法的合理性和可解释性。

4.4 科技论文新颖值与被引量分析

学术论文的新颖性与被引情况作为学术评价的两个重要维度，其之间的关系也受到一些学者的关注，如逯万辉等[12]研究发现主题新颖性较高的学术论文被引量通常要高于新颖性较低的文章。笔者也对文章的新颖值和被引量之间的关系进行了分析，结果如图 6 所示（方框内为被引量小于 1 500 的文章）。从图 6 可以看出，一方面高被引的文章具有高新颖值，如被引量大于 500 的文章新颖值均大于 0.6；另一方面新颖值较高的文章更容易产生高被引量，与逯老师等的实验结果相一致。关于科技论文的新颖值与被引量在本实验中没有呈现绝对的正相关关系的问题，笔者分析主要有两个原因：①文章的被引量随时间而增加，对于近两三年的文章而言，其真正被引量还没有体现出来；②科技论文的影响力不单单取决于文章新颖性大小，还会受到研究领域热度等因素的影响。

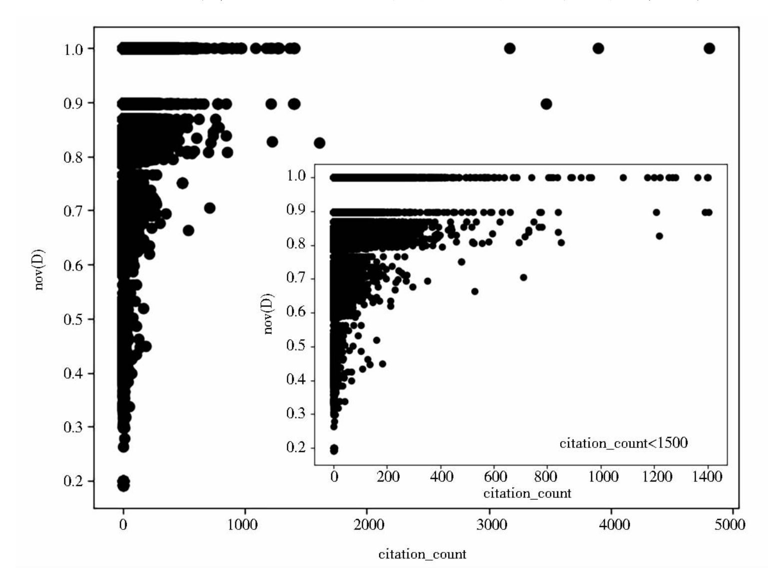


图 6 科技论文新颖值与被引量分布示意

5 总结与讨论

作为科技论文的核心要素，问题和方法的组合能够在一定程度上体现科技论文的新颖性。科技论文中的问题方法组合具体来说包括：新问题 + 新方法、新问题 + 老方法、老问题 + 新方法、老问题 + 老方法，而老问题 + 老方法又包含了老问题 + 老方法的老组合和老问题 + 老方法的新组合。受限于问题、方法词抽取困难，以往研究在计算学术论文新颖值时往往不对关键词的词汇功能加以区分，所以会掩盖老方法用于新问题、新方法用于老问题以及老问题 + 老方法新组合的新颖度。本文基于已有的科技论文问题方法抽取模型提出了一种基于问题 - 方法组合共现率计算科技论文新颖度的方法，分别计算科技论文的问题新颖度、方法新颖度、问题 - 方法组合新颖度以及论文整体新颖度。基于组合创新的思想，提出了 4 种创新类型以及根据新颖值识别文章所属创新类型的方法。最后，本文计算了 20 多万篇 ACM 论文的新颖值，并通过随机抽取的 10 篇科技论文计算结果分析证明了所提公式的合理性、可操作性以及计算结果的可解释性。

本研究还存在一定的局限性，提出的计算方法只考虑了科技论文的问题和方法而忽略了其他维度的新颖性，如新观点、新结论等，未来可以进一步扩展，例如文章的观点和结论更多地是以句子的形式出现在文章的摘要和首尾部分，因此需要先研究文章观点、结论句的识别，再进一步地将文章的观点、结论纳入计算公式，从更细粒度的层面更综合地度量论文的新颖性。

参考文献：

[1] 新华网. 习近平：关于《中共中央关于制定国民经济和社会发展第十四个五年规划和二〇三五年远景目标的建议》的说明[EB/OL]. [2021 - 06 - 01]. http://m. xinhuanet. com/2020-11/03/c_1126693341. htm.
[2]《科技查新教程》编写组. 科技查新教程[M]. 北京：机械工业出版社，2001.
[3] 尹新天. 中国专利法详解[M]. 北京：知识产权出版社，2011.
[4] 陈建青. 对我国学术论文创新性评审的几点思考[J]. 青年记者，2013(18)：33 - 35.
[5] NONAKA I. A dynamic theory of organizational knowledge creation[J]. Organization science，1994，5(1)：14 - 37.

[6] FLEMING L. Recombinant uncertainty in technological search[J]. Management science, 2001, 47(1):117－132.
[7] UZZI B, MUKHERJEE S, STRINGER M, et al. Atypical combinations and scientific impact[J]. Science, 2013, 342(6157):468－472.
[8] MISHRA S, TORVIK V I. Quantifying conceptual novelty in the biomedical literature[EB/OL]. [2021－03－01]. http://hdl.handle.net/2142/90328.
[9] 陆伟,李鹏程,张国标,等. 学术文本词汇功能识别——基于BERT向量化表示的关键词自动分类研究[J]. 情报学报,2020,39(12):1320－1329.
[10] 杨锋,梁樑,苟清龙,等. 同行评议制度缺陷的根源及完善机制[J]. 科学学研究,2008, 26(3):569－572.
[11] 沈律. 科技创新的一般均衡理论——关于科技成果创新度评价的科学计量学分析[J]. 科学学研究,2003,21(2):205－209.
[12] 逯万辉,苏金燕,余倩. 学术成果主题新颖性与学术引用的相关关系研究[J]. 情报资料工作,2018(6):68－73.
[13] 杨建林,钱玲飞. 基于关键词对逆文档频率的主题新颖度度量方法[J]. 情报理论与实践,2013,36(3):99－102.
[14] 任海英,王德营,王菲菲. 主题词组合新颖性与论文学术影响力的关系研究[J]. 图书情报工作, 2017,61(9):87－93.
[15] 杨京,王芳,白如江. 一种基于研究主题对比的单篇学术论文创新力评价方法[J]. 图书情报工作,2018,62(17):75－83.
[16] 许丹,徐爽,陈斯斯,等. 基于自然语言词对法的文献主题新颖性探测研究[J]. 图书情报工作, 2018, 62(8):130－138.
[17] 逯万辉,谭宗颖. 学术成果主题新颖性测度方法研究——基于Doc2Vec和HMM算法[J]. 数据分析与知识发现,2018,2(3):22－29.
[18] 钱玲飞,杨建林,邓三鸿. 人文社会科学学科创新力单指标评价[J]. 图书与情报,2013(2):93－98.
[19] 熊彼特. 经济发展理论[M]. 郭武军,吕阳,译. 北京:华夏出版社,2015.
[20] GARCIA R,CALANTONE R. A critical look at technological innovation typology and innovativeness terminology: a literature review[J]. Journal of product innovation management, 2002, 19(2):110－132.
[21] HENDERSON R M, CLARK K B. Architectural innovation: the reconfiguration of existing product technologies and the failure of established firms[J]. Administrative science quarterly, 1990, 35:9－30.
[22] BOWER J L, CHRISTENSEN C M. Disruptive technologies: catching the wave[J]. Harvard business review, 1995, 73(1):43－53.
[23] 宋子良. 基础研究创新分类初探[J]. 科技管理研究,1995,15(4):27－29.

作者贡献说明:
钱佳佳:设计研究方法,实验与数据分析,论文撰写;
罗卓然:提出研究方案,论文修改;
陆伟:确定研究方向,论文审阅。

Novelty Measurement and Innovation Type Identification of Scientific Literature Based on Question-Method Combination

Qian Jiajia　Luo Zhuoran　Lu Wei
School of Information Management Wuhan University, Wuhan 430072

Abstract: [**Purpose/significance**] Novelty measurement is an important part of scientific achievement evaluation. This paper aims to propose a method of novelty measurement and innovation type identification of scientific papers based on the combination of question and method. [**Method/process**] Based on the word frequency principle, this paper calculated the question novelty, method novelty and question-method combination novelty respectively, and then calculated the overall novelty of the paper by weight assignment. In addition, based on the theory of combination innovation, this study proposed four types of innovation from the perspective of scientific paper question-method combination and a method to identify the type of innovation according to the novelty value. [**Result/conclusion**] Finally, this paper conducts an empirical study based on more than 200,000 ACM papers from 1951 to 2018, and proves that the novelty measurement method and innovation category identification method proposed in this paper are scientific, reasonable and feasible.

Keywords: scientific literature　novelty measurement　combination innovation　question-method combination